2019 RISP International Workshop on Nonlinear Circuits, Communications and Signal Processing (NCSP2019)
Honolulu, Hawaii, USA, March 4-7, 2019

Study on Nonlinear Relationships between Semantic Primitives and Emotional Dimensions for Improving Three-layered Model

Xingyu Liu, Reda Elbarougy, Masato Akagi

School of Information Science, Japan Advanced Institute of Science and Technology
1-1 Asahidai, Nomi, Ishikawa, 932-1292, JAPAN
E-mail: {s1710233, akagi}@jaist.ac.jp, elbarougy@du.edu.eg

Abstract

Three-layered model is a perceptual model mimicking the process of human perception on speech emotion. However, previous studies based on the three-layered model only focused on the linear relationships in the emotion perception process among the three layers. In this research, nonlinear relationships between the second layer (semantic primitives) and the first layer (emotion dimensions), the top two layers of the three-layered model, were investigated by taking advantage of fuzzy inference system (FIS) as an estimator. Effective semantic primitives to describe human perception on emotion dimensions were selected from 28 semantic primitive candidates. Evaluation results show that semantic primitives selected by the proposed method are effective to describe emotion dimensions and can be used to construct an improved three-layered model.

1. Introduction

Speech is one of effective media to express emotions. Clarifying the mechanism of human speech emotion perception and building a model to mimic this mechanism has become an important research topic in the field of human computer interaction and affective computing. Emotions are usually defined in tow approaches, the categorical approach such as joy, sadness, anger and so on and dimensional approach spanned by Activation and Valence [1]. Although dimensional approach can represent degrees of emotional states comparing with categorical approach, accurately mapping of acoustic features to emotion dimensions is still challenging.

Based on the assumption that human perception of speech emotion is a multiple-layer process [2], a three-layered model as shown in Figure 1 was proposed by Huang and Akagi [3] to describe human perception with the categorical approach. The model was consisted of acoustic features as the bottom layer, semantic primitives that represent human feelings when hearing expressive speech as the middle layer, and emotion categories as the top layer. Elbarougy [4] revised the model with the dimensional approach using Activation, Valence and Dominance, and achieved a speech emotion recognition system using the same semantic primitives as Huang. The semantic primitives in these research were chosen by calculating correlation coefficient (CC) between emotion states and semantic primitives, ignoring the insignificance of CC in representing non-linear relationships. In addition, semantic primitives utilized in these research were selected under the categorical approach, it is not clear whether they are suitable for describing emotion dimensions . Therefore, discussing on the non-linear relationships between semantic primitives and emotion dimensions (especially for Activation and Valence) becomes necessary.

This research aims to investigate non-linear relationships between semantic primitives and emotion dimensions, and to select appropriate semantic primitives. To achieve these goals, fuzzy inference system (FIS) was used as an estimator to estimate the relationships between semantic primitives and emotion dimensions. Then the effect of different relationships on predicting emotion dimension was discussed. Semantic primitives related to emotion dimensions were selected based on the discussion. Finally, the effectiveness of selected semantic primitives was evaluated by comparing with those utilized in previous research.

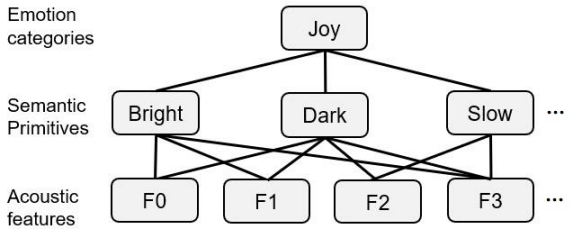

Figure 1: Structure of a three-layered model

2. Listening test

The aim of the listening test is to collect subjects' evaluation scores on each semantic primitive when hearing affective speech. The listening test is a preparatory work for training FIS.

Table 1: 28 candidates for semantic primitives and PCC

SP	max	min	average	SP	max	min	average
1.Bright	0.90	0.56	0.77	15.Sharp	0.95	0.64	0.81
2.Dark	0.93	0.63	0.81	16.Fast	0.88	0.77	0.84
3.High	0.91	0.69	0.83	17.Slow	0.87	0.73	0.84
4.Low	0.86	0.73	0.82	18.Definite	0.87	0.58	0.74
5.Strong	0.95	0.77	0.87	19.Warm	0.86	0.42	0.71
6.Weak	0.93	0.70	0.86	20.Soft	0.88	0.55	0.76
7.Calm	0.89	0.63	0.77	21.Powerful	0.90	0.75	0.84
8.Unstable	0.88	0.52	0.78	22.Roundish	0.73	0.44	0.65
9.Well-modulated	0.93	0.57	0.78	23.Violent	0.89	0.68	0.83
10.Monotonous	0.88	0.56	0.77	24.Hard	0.89	0.65	0.82
11.Heavy	0.94	0.39	0.76	25.Fluent	0.84	0.56	0.74
12.Clear	0.93	0.32	0.75	26.Light	0.90	0.53	0.74
13.Noisy	0.95	0.57	0.84	27.Thin	0.78	0.31	0.64
14.Quiet	0.92	0.69	0.86	28.Thick	0.86	0.35	0.69

2.1 Stimuli and subjects

CASIA, a Chinese emotion speech dataset made by Chinese Academy of Science, was used for the listening test. 208 Chinese utterances were selected from the dataset as stimuli which included four different emotion states: joy, sadness, anger and neutral. Activation and Valence scores of these utterances had already been annotated.

Five male and five female native Chinese speakers whose age ranged from 24 to 29 years old were asked to participate in the test. Before the test, a pre-practice was done to help subjects to understand the meaning of each semantic primitive. Utterances for pre-practice were selected by another group of Chinese native speakers.

2.2 Candidates for semantic primitives

This research used 28 adjectives, shown in Table 1, as candidates of semantic primitives because these adjectives were proved to be used frequently in describing speech emotion by Zhang and Akagi [5] . The candidates from No.1 to No.17 were also used for the middle layer as semantic primitives in previous research [3] and [4].

2.3 Procedure

For each candidate, subjects were asked to listen to the 208 stimuli to evaluate the degree of this candidate using a 5-point scale approach: 1-Dode not feel at all; 2-Seldom feels; 3-Feels a little; 4-Feels; 5-Feels very much. The subjects evaluated all 28 candidates with 208 utterances totally.

2.4 Inter-rater agreement evaluation

After evaluating all 28 adjectives, the agreement between subjects to verify whether subjects have similar perception on each candidate was carried out. It is because we clarify whether the results of the listening test can used to train FIS. Low agreements between subjects could produce a lot of noise in training data and affect the performance of FIS.

Pearson correlation coefficient (PCC) was used to verify the agreement between subjects. Table 1 shows the PCC evaluation results of 28 candidates. Most of candidates for their PCC were above 0.7. This result indicates that the subjects had highly consistent perception. Therefore, the data collected from the listening test were usable for training FIS.

3. Evaluation

3.1 FIS training

FIS is a human-knowledge-based mathematical estimator and works well on non-linear functional modeling. For this reason, it is appropriate to be used as an estimator to investigate the non-linear relationships between semantic primitives and emotion dimensions. Particularly, a revised FIS called ANFIS [6] was used in this research because ANFIS has the ability to generate fuzzy rules automatically.

The evaluation score of each candidate of semantic primitive collected from the listening test was one input of ANFIS, the value of Valence or Activation was the output of ANFIS. Therefore, ANFIS contained 28 inputs and one output. Since 208 utterances were used in the listening test, the size of the training data was 208. The data was divided into ten disjoint subsets to train the FIS using ten-cross validation method. Each subset included all the four different emotion categories. When training the FIS, nine subsets combined the input training set, and one subset was testing set. The data was also

normalized using the max-min normalization approach:

$$z_i = \frac{x_{i,n} - min(x_i)}{max(x_i) - min(x_i)} \quad (1)$$

where i is the ith semantic primitive, n is the nth utterance and x is the evaluation score.

Original data ranged from 1 to 5 and after normalization they ranged from 0 to 1.

3.2 Estimation and grouping of relationships

By using the trained ANFIS, the relationship between each semantic primitive and emotion dimension was estimated in this subsection. In addition, in order to discuss the effect of these relationships on the prediction of emotion dimensions, they were divided into several different patterns.

The method of evaluating relationships between candidates of semantic primitives and emotion dimensions was to fix the other 27 inputs as constant 0.5, change the remaining one input from 0 to 1, then observe the outputs of ANFIS.

The relationships between input and output of ANFIS were grouped in three patterns:

- pattern 1- monotonically increasing (or decreasing)
- pattern 2- U-shape or n-shape
- pattern 3- neither belong to Pattern 1 nor 2

Figures 1 to 3 are examples to show these three different patterns and Table 2 shows all the grouping results. Candidates of semantic primitives are represented by their numbers in Table 1.

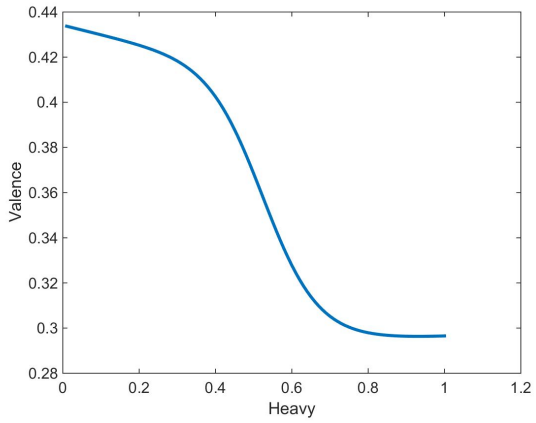

Figure 2: example for Pattern 1- the relationship between Heavy(No.11) and Valence

Although some candidates belong to a certain pattern, there showed small change in output value (the range of output changing were below 0.1) when these candidates of semantic primitives changed. These candidates were taken as not

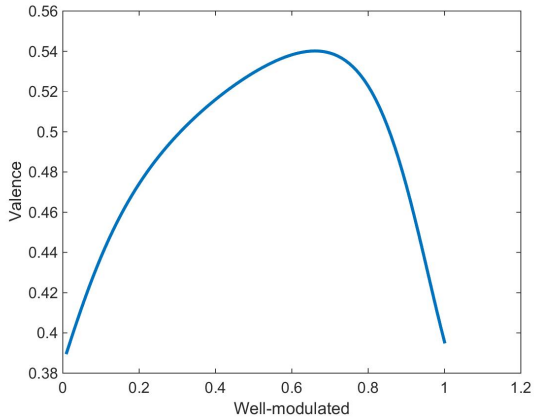

Figure 3: example for Pattern 2- the relationship between Well-modulated(No.9) and Valence

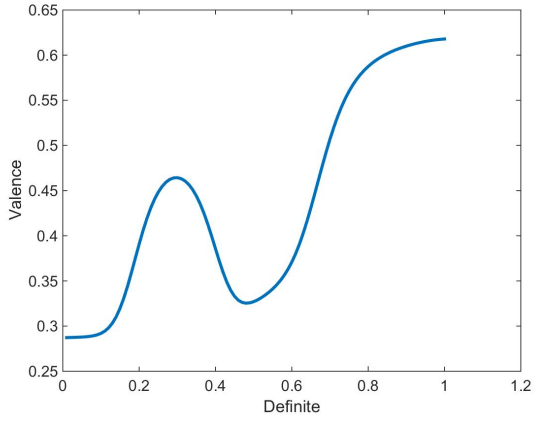

Figure 4: example for Pattern 3- the relationship between Definite(No.18) and Valence

Table 2: Grouping results. The numbers of strongly related semantic primitives were colored in red

	Pattern 1	Pattern 2	Pattern 3
Valence	1, 4, 5, 11, 12, 13, 14, 16, 17, 20, 21, 22, 25, 26	2, 3, 6, 7, 8, 9, 15, 23, 24, 27, 28	10,18, 19
Activation	1, 2, 3, 4, 5, 8, 9, 12, 14, 15, 16, 17, 21, 23, 24, 25	6, 7, 10, 11, 13, 18, 19, 20, 22, 26, 28	27

related to emotion dimensions and were excluded from each pattern. Strongly related candidates of semantic primitives were colored in red as shown in Table 2.

3.3 Effect on predicting emotion dimensions

To clarify how these patterns of relationships affect the accuracy of emotion dimension prediction, other sets of ANFIS were trained and tested. The input of ANFIS were (1) candidates of semantic primitives belonging to the same pattern,(2) combine candidates of semantic primitives belonging to different patterns together. In addition, original 17 semantic primitives selected in [3] were also tested. For the evaluation, distances (root mean square error, RMSE) between humans' evaluation scores and estimates by the trained ANFIS were used. Figure 4 shows the RMSEs for Valence and Figure 5 shows RMSEs for Activation.

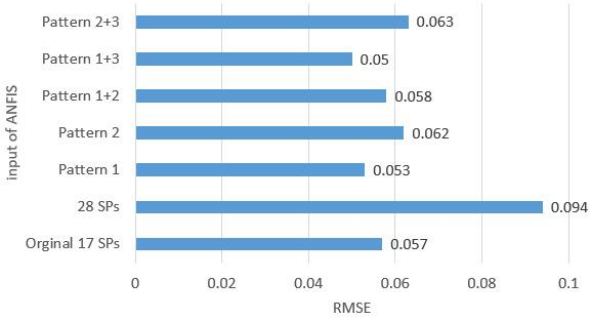

Figure 5: RMSEs for Valence

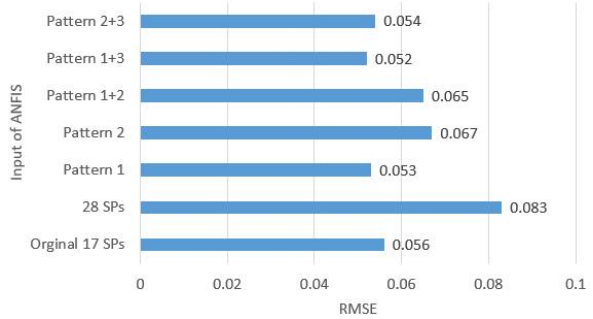

Figure 6: RMSEs for Activation

3.4 Discussion

From Figures 4 and 5, it can be clearly found that combining candidates of semantic primitives belonging to Pattern 1 and Pattern 3 achieved the best emotion dimension prediction results in both Valence and Activation dimensions. From this result, semantic primitives that highly related to Valence and Activation were selected. Valence related semantic primitives are: Bright, Heavy, Clear, Soft, Powerful, Roundish, Fluent, Light and Warm. Activation related semantic primitives are: Bright, Dark, High, Low, Strong, Unstable, Well-modulated, Quiet, Sharp, Fast, Powerful, Violent, Fluent and Thin.

In addition, the prediction accuracy seems not to be directly related to the number of input semantic primitives because the most semantic primitives produced the worst prediction accuracy. Other than this, the performance of original 17 semantic primitives was poor than those belonging to Pattern 1 or Pattern 1+3. This also proves that semantic primitives related to emotion categories are not necessarily related to emotion dimensions.

It should be pointed out that which pattern the semantic primitive belongs to is not an absolute conclusion. When changing training data and testing data, relationships between some semantic primitives and emotion dimensions may change so that these semantic primitives belong to different patterns in such condition.

4. Conclusions

The results of this research showed that relationships between many semantic primitives and emotion dimensions are non-linear and it is important to take such relationships into account when selecting effective semantic primitives. Semantic primitives selected by approach described in this paper produced better prediction accuracy than previous selecting method (CC). This finding may provide reference for improving the performance of the three-layered model in speech emotion recognition or synthesis.

References

[1] Russell, J. A., "Core affect and the psychological construction of emotion," Psychological Review, 110(1), pp.145-172. 2003.

[2] K. R. Scherer, "Personality Inference from Voice Quality: The Loud Voice of Extroversion," European Journal of Social Psychology, 8, 467–487, 1978.

[3] C. Huang, and M. Akagi, "A three-layered model for expressive speech perception," Speech Communication, 50(10), pp.810-828, 2008.

[4] R. Elbarougy, and M. Akagi. "Improving speech emotion dimensions estimation using a three-layer model of human perception," Acoustical Science and Technology 35.2, pp.86- 98, 2014.

[5] C. Zhang and M. Akagi, "Study on differences between perceptions of Japanese and Chinese emotional speech by Japanese and Chinese listeners," 2018 International Workshop on Nonlinear Circuits, Communications and Signal Processing (NCSP'18), Hawaii, USA, pp.359-362, 2018.

[6] J.-S.R. Jang, "ANFIS: adaptive-network-based fuzzy inference system," IEEE Transactions on Systems, Man, and Cybernetics,23(3), pp.665-685, 1993.